DOCUMENT RESUME

ED 350 335 TM 019 094

AUTHOR	Valley, John R.
TITLE	The SAT: Four Major Modifications of the 1970-85 Era.
INSTITUTION	College Entrance Examination Board, New York, N.Y.
REPORT NO	CBR-92-1
PUB DATE	92
NOTE	35p.
AVAILABLE FROM	College Board Publications, Box 886, New York, NY 10101-0886 ($12).
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Achievement Tests; *College Entrance Examinations; Educational History; Equated Scores; High Schools; High School Students; *Item Response Theory; Psychometrics; Standardized Tests; *Test Construction; Test Content; Test Format; Test Reliability; Test Validity; Writing Evaluation
IDENTIFIERS	*Scholastic Aptitude Test; Test of Standard Written English; *Test Revision

ABSTRACT

From 1970 to 1985, the Scholastic Aptitude Test (SAT) underwent major modifications caused by: (1) the addition of the Test of Standard Written English (TSWE) to the College Board's Admissions Testing Program (ATP); (2) the passage of test disclosure legislation; (3) the institution of test sensitivity reviews; and (4) the use of item response theory equating in SAT scores. This report discusses these modifications as they relate to the SAT's content, format, development procedures, psychometric characteristics, and statistical procedures; and it concludes that despite these major modifications and other concomitant minor changes, the SAT has maintained its stability and continuity. Two figures and eight tables present study findings, and there is an appendix of SAT item-type examples. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED350335

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

ROBERT G. CAMERON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

College Board Report

No. 92-1

The SAT®: Four Major Modifications of the 1970–85 Era

John R. Valley

BEST COPY AVAILABLE

The SAT®: Four Major Modifications of the 1970–85 Era

John R. Valley

College Board Report No. 92-1

College Entrance Examination Board, New York, 1992

John R. Valley is an Educational Consultant.

Acknowledgments

The author acknowledges the useful comments and excellent suggestions provided by Caroline Crone and Gary Marco of Educational Testing Service, who reviewed an earlier version of this report.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,800 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $12.

All figures and tabular material in this report are reprinted by permission of Educational Testing Service, the copyright owner, unless noted otherwise.

Copyright © 1992 by College Entrance Examination Board. All rights reserved.
College Board, Scholastic Aptitude Test, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board.
Printed in the United States of America.

CONTENTS

Abstract 1

Introduction 1

Background 2
- Dynamics of the SAT 2
- General Description of the SAT 3
- The Pre-October 1974 SAT 7

Modifications to the SAT After TSWE 8

Observations Pertinent to the Introduction of TSWE 10
- Test Difficulty Specifications 10
- Biserial Correlations 12
- Score Conversions 12
- Relative Test Difficulty 12
- Speededness 12
- Reliability 13
- Correlational Patterns 13

Modifications to Accommodate Test Disclosure Legislation 14

Observations Regarding Test Disclosure 15

Modifications Related to Test Sensitivity Review 15
- The Rationale 15
- The Critical Elements 16
- The Process 16

Observations Regarding Test Sensitivity Review 16
- Minority-Relevant Reading Passages 16
- Gender References 16

Use of Item-Response Theory Equating 17
- Linear Equating 17
- IRT Equating 18

Observations Related to IRT Equating 19
- Equating Indices 19
- Appropriateness of Equating Methods 19
- Scale Stability 20
- Results 20

Summary 21

References 21

Appendix: SAT Item-Type Examples 22

Figures

1. Item-order specifications for SAT-verbal sections 3

2. Item-order specifications for SAT-mathematical sections 6

Tables

1. Specified and Actual Numbers of Items within Various Classifications for SAT-Verbal Forms 4

2. Specified and Actual Numbers of Items within Various Classifications for SAT-Mathematical Forms 5

3. Statistical Specifications for SAT-Verbal and SAT-Mathematical Forms from 1966 to the Present 6

4. Specified and Actual Item Statistics for November and December SAT-Verbal Forms from 1970 to 1984 7

5. Specified and Actual Item Statistics for November and December SAT-Mathematical Forms from 1970 to 1984 7

6. Summary of Changes Made to SAT Item Types, Content, and Test Format from March 1973 to January 1985 8

7. Summary of Statistical Characteristics of November and December SAT Forms from 1970 to 1984 10

8. Equating Methods Used for November and December SAT-Verbal and SAT-Mathematical Equatings from 1970 to 1984 18

ABSTRACT

From 1970 to 1985, the Scholastic Aptitude Test (SAT) underwent major modifications caused by (1) the addition of the Test of Standard Written English (TSWE) to the College Board's Admissions Testing Program (ATP), (2) the passage of test disclosure legislation, (3) the institution of test sensitivity reviews, and (4) the use of item-response theory equating in SAT scores. This report discusses these modifications as they relate to the SAT's content, format, development procedures, psychometric characteristics, and statistical procedures, and concludes that despite these major modifications and other concomitant minor changes, the SAT has maintained its stability and continuity.

INTRODUCTION

Today in many circles the College Board and the Scholastic Aptitude Test (SAT) are synonymous. However, the founding of the College Board, the association of secondary schools and colleges that sponsors the SAT, predates the test by almost a quarter of a century. June 1926 marks the first administration of the SAT under College Board sponsorship, when over 8,000 students were tested.

The basic structure of the SAT, as we know it today, was established over the next 15 years. The following are significant landmarks of the era:

- In 1929 the test was divided into two sections, verbal and mathematical.
- In 1937 a second annual administration of the test was offered for the first time.
- In 1941 the SAT's score scale was permanently established by using the 11,000 students tested at the April administration as the standardization group.
- In 1941, at the June administration, equating of scores was begun as a routine procedure that has been followed continuously ever since.

Since 1941 the SAT has been a multiform measure of developed verbal and mathematical ability yielding scores that are equated with the 1941 standard and reported on a 200–800 scale (with a mean of 500 and a standard deviation of 100).

In the 1970–85 era, four major modifications to the SAT took place. These modifications merit being designated as major because there was public awareness of them, as test takers, schools, and colleges had to be advised of the change; or the modification involved the adoption of a new policy; or the modification involved a shift in what is undoubtedly one of the most sensitive procedures of the entire testing program, score equating. These four modifications were:

1. *Modifications to accommodate the Test of Standard Written English (TWSE).* In the fall of 1974 the TSWE was introduced experimentally as part of the College Board's Admissions Testing Program (ATP). As a result the SAT itself was shortened from 3 to 2½ hours. This was accomplished by replacing an old item type with a new one in the mathematical sections and redistributing the number of certain item types in the verbal sections. The TSWE became a regular fixture of the ATP in 1977, and fine-tuning modifications were made to the SAT as late as 1980.
2. *Modifications to accommodate disclosure legislation.* Quite likely test takers, schools, and colleges were even more aware of this modification than of the first. In response to disclosure legislation passed by New York State, the College Board began the postadministration publication of complete SATs in October 1980. The legislation also required that answers accepted as correct be made known. Test takers could request detailed reports explaining how their test scores were derived.
3. *Modifications in response to sensitivity reviews.* The third modification was the requirement that the SAT undergo minority, cultural, and gender sensitivity reviews. Formal policies and procedures developed by Educational Testing Service (ETS), which develops and administers the SAT for the College Board, required that the SAT, along with other tests, undergo sensitivity reviews as part of normal test development and production. Implementation of the change was steady and continuous over the 1970–85 period.
4. *Modification of operational equating.* The fourth modification was confined to internal processing. It was associated with a change in the difficulty specifications for SAT-verbal sections and involved the operational use of item-response theory (IRT) equating for the first time in 1982. This equating method is more sophisticated than those previously used. It uses information about the performance of individual test items and provides a more accurate equating of a test form that is built to new statistical specifications. Since it is an operational procedure employed after the test has been administered, test takers were unaware of the change.

Many of these modifications are described in two College Board technical manuals for the SAT and Achievement Tests (Angoff 1971; Donlon 1984). Recently ETS published *Trends in SAT Content and Statistical Characteristics and Their Relationship to SAT Predictive Validity* (Marco et al. 1990). As the title implies, the study relates changes in the SAT and its statistical specifications to observations about the test's predictive validity in the 1970–85 era. Using that recent report as the primary source of information, this report documents the major modifications in the test during the 1970–85 period, without reference to the test's predictive validity, and provides more detailed information about these modifications than is available in the 1984 technical handbook.

A somewhat arbitrary distinction is made in this report between "planned" changes and modifications to the SAT

and "observed" changes. Planned changes are limited to specific deliberate actions taken by the College Board or by ETS. These actions involve the structure of the test, information about it, its content, timing, statistical specifications, and so on. These actions also deal with matters of processing, changes in test development procedures, or equating, for example.

Some planned changes or modifications might be readily noticed by test takers, such as a reduction in the total test time, expanded information about the test, or information describing a change in the types of questions a candidate might expect. Other planned changes may deal with matters beyond the ken of test takers, such as changes in processing and policies related to test development, quality control and review procedures, equating, content and test specifications, and so on.

Planned changes or modifications bring consequences or second-order changes that are significant and merit notice. Information regarding such changes is derived from test analyses routinely conducted after each administration, from findings of program research studies, or from reviews of program operations. Data or information of these sorts are collectively considered observations about changes.

To illustrate further, assume a decision is made to reduce test difficulty by a specified amount, and steps are taken to implement the decision. The decision and supporting actions are planned changes or modifications. At a later time, data are gathered to ascertain the extent to which the change had been implemented successfully. The data may show that the general objective was accomplished, i.e., test difficulty was reduced, but the precise target level was not attained. This report would categorize the attained difficulty level as an observation about a modification. This distinction is followed in the organization of this report, which first discusses the steps taken to modify the SAT and then reports observations based on data pertinent to the modification.

Changes in the abilities of students taking the test are not discussed beyond noting here the general decline observed for the period 1971–85. Data for college-bound seniors show that both SAT-verbal and SAT-mathematical scores declined for more than a decade, followed by a slight upward trend in the 1982–85 period. Despite this increase in scores, by 1985 average SAT scores had not returned to their 1971 levels.

Before dealing with each of the four major modifications, it is helpful to briefly describe the SAT itself.

BACKGROUND

Dynamics of the SAT

The SAT must be viewed not as a stand-alone measure but as the keystone of the College Board's Admissions Testing Program. Along with the SAT, the ATP now consists of:

1. The Student Descriptive Questionnaire (SDQ), completed by candidates at home and turned in at the test center.
2. The TSWE, administered with the SAT as part of a morning program of tests.
3. Fifteen Achievement Tests dealing with specific subject areas commonly taught in secondary schools.

The ATP facilitates students' transition from secondary schools to colleges and universities. It does so by supplementing the information schools provide about candidates for admission as well as adding to the information and impressions candidates furnish about themselves on their admission applications and in admission interviews. Consequently, the ATP is sensitive to the information requirements of both school and college officials, particularly as their needs are perceived to change over time.

The interactive relationship of the SAT, as part of the ATP, is seen quite readily in connection with the addition of the TSWE. In the early 1970s many colleges and universities indicated a need to know the extent of applicants' understanding of basic conventions of written English in order to place students in appropriate English courses. Consequently the SAT was modified and the TSWE added to the ATP.

One must also note that the SAT is more than simply a measure of individual student development. Colleges and universities have found it useful to aggregate the test results of their applicants annually and to compare these statistics with those for their admitted students as well as their enrolled freshman classes. Institutions conduct such analyses for purposes of year-to-year comparisons. At times these data may be shared with other institutions, the public at large, or secondary schools. At the institutional level, then, year-to-year aggregation of SAT data provides one measure of the quality of student input to an institution and permits trends in these data to be observed and recognized.

In the 1930s only a few thousand students, representing a very limited number of high schools, took the SAT. Except to the relatively few students involved and the colleges to which scores were reported, there was no particular import attached to the results. Compare those years to the current situation! Now close to a million high school students take the test each year. SAT results are watched as one of the indicators of the health of the nation's education. In the 1970s, the observation that average SAT scores for college-bound students were declining was cause for concern in many quarters. Despite cautions against overinterpreting small fluctuations in test scores, the general public's reaction to that observation signaled that the SAT was being regarded as a kind of national education barometer.

A shortened version of the test, the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT), affords students the opportunity to experience a test similar in content and format to the SAT before applying for admission to college. Its scores play a useful role in advising college-bound students. PSAT/NMSQT results also serve to qualify students for the National Merit Scholarship program. In 1985, there were 1,185,571 students who took the PSAT/NMSQT.

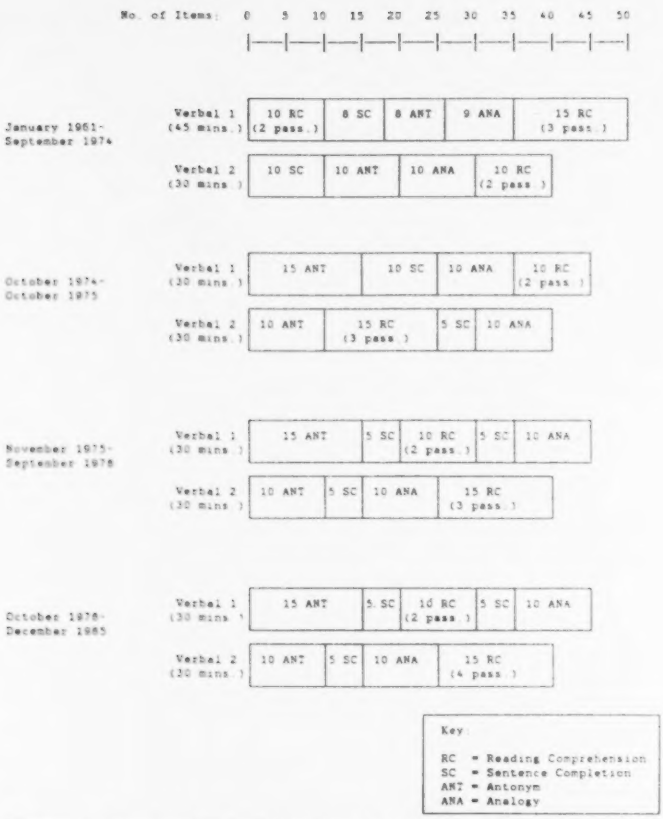


Figure 1. Item-order specifications for SAT-verbal sections.

The preceding discussion suggests that the SAT is influenced by two opposing tensions. One is the need for stability and continuity to maintain an interpretive linkage over time for education institutions and agencies that wish to apply data to individual and collective decisions and issues. The second, representing the need for change, derives its influence from educational, cultural, social, and technological developments in our society that collectively transform the environment in which the test is expected to function. Three examples serve to illustrate these influences for change.

First, in the 1970s, increased societal sensitivity to ethnic, cultural, social, and gender biases not only changed the processes by which the SAT was developed but also modified the analyses of the results of the test. Second, test disclosure legislation led to increased test item production and test development staff expansion, increased pretesting, increased final form production, and substantial changes in the kind and amount of information provided to candidates both before and after test administration. Third, technological advances such as high-speed electronic computers and data-processing equipment have held down costs and allowed reporting schedules to be maintained despite increased volume. Meanwhile, theoretical psychometric developments supported by the availability of high-speed data-processing technology found practical application for item-response theory in the task of equating SAT scores. These examples are all discussed in greater detail in the following sections.

General Description of the SAT

The development and production of each new edition of the SAT is controlled by very detailed item type, content, and statistical specifications. The SAT-verbal sections use four types of items: reading comprehension, sentence completion, antonyms, and analogies (see Figure 1). For reading comprehension items there is a specified distribution of content covering biological science, social studies, humanities, narrative, argumentative, and synthesis. Reading comprehension items are classified additionally by skills: main idea, supporting idea, inference, application, evaluation of logic, style, and tone. The other three item types—antonyms, analogies, and sentence completion—are assigned to content categories that are different from those used for reading comprehension items. Those categories are aesthetic-philosophical, world of practical affairs, science, and human relationships. (See Table 1.) Additionally content subsets are used for each discrete item type. Sentence completion items are separated by structure (one or two missing words). Structure is used also, but with a different connotation, in connection with antonyms; there structure

9 BEST COPY AVAILABLE

Table 1. Specified* and Actual Numbers of Items within Various Classifications for SAT-Verbal Forms**

Item Type Classification	*Specified (Jan. 1961–Sept. 1974)*	*Nov. & Dec. Actual (1970–1973)*	*Specified (Oct. 1974–Sept. 1978)*	*Nov. & Dec. Actual (1974–1977)*	*Specified (Oct. 1978–Present)*	*Nov. & Dec. Actual (1978–1984)*
Sentence Completions						
Content						
Aesthetics/philosophy	4	4–5	4	3–5	4	4–4
World of practical affairs	5	5–6	4	4–5	4	4–5
Science	5	5–5	4	3–4	4	3–4
Human relationships	4	3–4	3	3–3	3	3–3
(Total)	(18)	(18)	(15)	(15)	(15)	(15)
Antonyms						
Content						
Aesthetics/philosophy	4	3–5	6	5–6	6	4–7
World of practical affairs	5	4–7	6	6–7	6	5–7
Science	5	4–5	7	6–7	7	4–8
Human relationships	4	3–5	6	5–6	6	5–9
(Total)	(18)	(18)	(25)	(25)	(25)	(25)
Analogies						
Content						
Aesthetics/philosophy	5	4–6	5	5–5	5	4–6
World of practical affairs	5	4–6	5	5–6	5	5–7
Science	5	5–6	5	4–5	5	4–6
Human relationships	4	3–5	5	4–5	5	4–6
(Total)	(19)	(19)	(20)	(20)	(20)	(20)
Reading Comprehension						
Content						
Narrative	5	5–5	5	5–5	2–5	3–5
Biological science	5	0–10	0–5***	5–5	2–5	2–5
Physical science	5	0–5	0–5***	0–0	2–5	3–5
Argumentative	5	5–10	5	5–5	2–5	3–5
Humanities	5	5–5	5	5–5	2–5	3–5
Synthesis	5	0–5	0	0–0	0	0–0
Social studies	5	5–5	5	5–5	2–5	3–5
Functional Skill						
Main idea	7	2–8	5	3–6	5	1–6
Supporting idea	7	4–10	5	4–10	5	2–8
Inference	12	11–16	9	6–10	9	8–11
Application	3	1–4	2	1–3	2	1–3
Evaluation of logic	3	1–6	2	1–3	2	1–4
Style and tone	3	2–5	2	2–2	2	0–3
(Total)	(35)	(35)	(25)	(25)	(25)	(25)

*The specifications applied to any new test administered during the indicated period.

**Expressed as ranges.

***Only one science passage was permitted on the test.

refers to single words versus phrases. Antonyms are also classified by parts of speech: verbs, nouns, and adjectives. Analogies are sorted by abstraction of terms (concrete, abstract, mixed) and independence of stem and key (independent or overlapping).

Three item types have been used in the SAT-mathematical sections (see Figure 2 and Table 2). Before October 1974 these sections consisted of regular mathematics and data sufficiency items. Since then regular mathematics items are still used, but four-choice quantitative comparison items have replaced five-choice data sufficiency items.

All SAT-mathematical items are further classified as arithmetic, algebra, geometry, or miscellaneous. Additionally, the item settings are classified as concrete or abstract. Finally SAT-mathematical items are classified according to ability level:

- Level 0—Recall of factual knowledge.
- Level 1—Perform mathematical manipulations.
- Level 2—Solve routine problems.
- Level 3—Demonstrate competence in mathematical ideas and concepts.

Table 2. Specified* and Actual Numbers of Items within Various Classifications for SAT-Mathematical Forms**

Item Type Classification	*Specified (Nov. 1969–Sept. 1974)*	*Nov. & Dec. Actual (1970–1973)*	*Specified (Oct. 1974–Dec. 1975)*	*Nov. & Dec. Actual (1974–1975)*	*Specified (Jan. 1976–Sept. 1981)*	*Nov. & Dec. Actual (1976–1980)*	*Specified (Oct. 1981–Present)*	*Nov. & Dec. Actual (1981–1984)*
Regular Mathematics								
Arithmetic	13	13–13	12–13	12–12	12–13	12–13	12–13	12–13
Algebra	11	11–13	11	11–11	11	10–11	11	11–11
Geometry	13	12–13	11	11–12	11	11–11	11	11–11
Miscellaneous	5	4–5	5–6	5–6	5–6	5–6	5–6	5–6
(Total)	(42)	(42)	(40)	(40)	(40)	(40)	(40)	(40)
Data Sufficiency								
Arithmetic	4–5	4–5						
Algebra	4–5	3–5						
Geometry	6–7	5–8						
Miscellaneous	3–4	3–4						
(Total)	(18)	(18)						
Quantitative Comparisons								
Arithmetic			6	6–7	6	6–7	6	6–6
Algebra			6	5–6	6	5–6	6	6–6
Geometry			5–6	5–6	5–6	5–6	5–6	5–6
Miscellaneous			2–3	2–3	2–3	1–3	2–3	2–3
(Total)			(20)	(20)	(20)	(20)	(20)	(20)
All								
Setting								
Concrete	11–31	12–24	11–31	13–16	11–31	10–16	11–21	14–19
Abstract	29–49	36–48	29–49	44–47	29–49	44–50	39–49	41–46
Ability								
Solve routine problems	0–8	7–10	12–21	14–20	2–21	10–18	0–21	11–17
Demonstrate comprehension of math ideas and concepts	22–30	21–26	22–31	22–25	22–41	23–32	22–43	24–31
Apply "higher" mental processes to math	30–38	26–31	17–26	16–24	17–36	16–22	17–38	17–21
(Total)	(60)	(60)	(60)	(60)	(60)	(60)	(60)	(60)

*The specifications applied to any new test administered during the indicated period.
**Expressed as ranges.

- Level 4—Solve nonroutine problems requiring insight or ingenuity.
- Level 5—Apply "higher" mental processes to mathematics.

Statistical specifications for the SAT, expressed in terms of delta distributions and biserial correlations, control distributions of item difficulties and the correlation of items with the total test score (see Table 3). For the SAT, the percentage of test takers who answer an item correctly is computed by dividing the number obtaining the correct answer by the number who reached the item. ETS uses a transformation of this percentage (referred to as delta) as the primary measure of item difficulty. The transformation is a normal deviate with a mean of 13 and a standard deviation of 4. Deltas bear an inverse relationship to the proportion correct. Before they are used in the test development process, raw deltas (observed on pretest samples of test takers) are equated and expressed on a delta scale based on a common reference population. This procedure takes into account differences in the ability levels of the analysis groups from which the data for the observed deltas were obtained.

The biserial correlation, which is derived from the correlation of item response (right versus wrong) with total test score, provides an index of the ability of the item to discriminate between high- and low-ability test takers. Biserial correlations together with deltas are used to maintain the test's statistical specifications (see Tables 4 and 5).

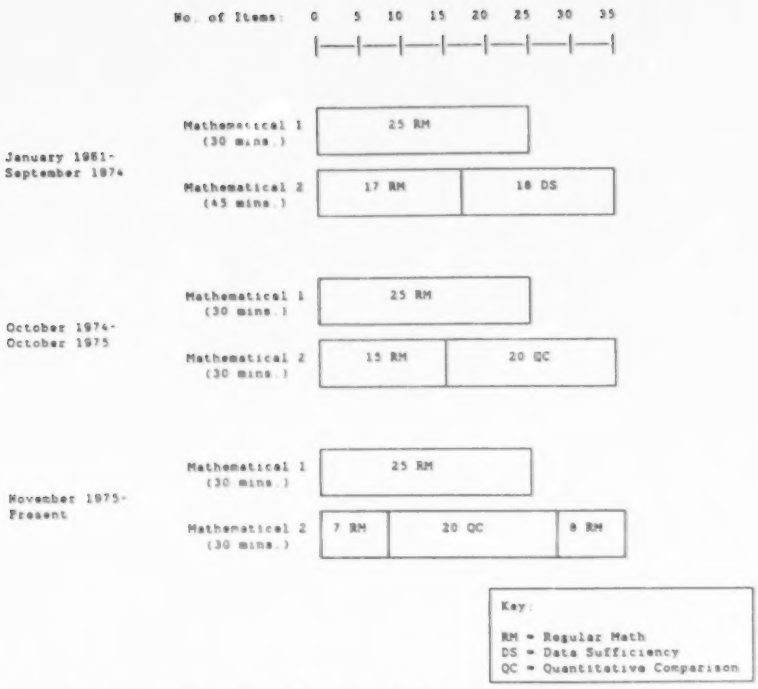


Figure 2. Item-order specifications for SAT-mathematical sections.

Table 3. Statistical Specifications for SAT-Verbal and SAT-Mathematical Forms from 1966 to the Present[a]

Item Difficulty (Equated Delta)	*SAT-Verbal*			*SAT-Mathematical*	
	Aug. 1966–Sept. 1974[b]	*Oct. 1974–Jan. 1982*[c]	*Jan. 1982–Present*	*Aug. 1966–Sept. 1974*	*Oct. 1974–Present*
≥ 18	0	0	0	3	3
17	2	2	0	4	4
16	4	4	2	4	4
15	8	10	6	4	4
14	10	10	14	5	4
13	10	6	10	5	4
12	10	6	8	5	4
11	10	6	7	8	8
10	10	8	7	8	8
9	8	8	10	7	8
8	7	10	8	4	5
7	6	8	6	2	1
6	3	4	4	1	2
≤5	2	3	3	0	1
Number of Items	90	85	85	60	60
Mean Delta	11.7	11.4	11.4	12.5	12.17–12.27
SD Delta	2.9	3.3	3.0	3.1	3.1–3.3
Mean Biserial r[d]	.42 (.47)	.43 (.48)	.41–.45 (.46–.50)	.47 (.53)	.47 (.53)

[a]The statistical specifications applied to any new form administered during the indicated periods.

[b]From August 1966 to July 1967 the statistical specifications for SAT-V were as follows: mean delta = 11.8, SD delta = 3.0, mean biserial r = .42.

[c]One of the two January 1982 forms was assembled to the specifications for the 1974–81 period.

[d]The mean biserial r is specified in terms of pretest items, which are not included in the total-score criterion. The equivalent means for a total-score criterion that includes the item, given in parentheses, are .05 higher for the SAT-V and .06 higher for the SAT-M.

Table 4. Specified and Actual Item Statistics for November and December SAT-Verbal Forms from 1970 to 1984

	Specified			*November Actual*			*December Actual*		
Year	*Mean Equated Delta*	*SD Equated Delta*	*Mean Biserial r*[a]	*Mean Equated Delta*	*SD Equated Delta*	*Mean Biserial r*	*Mean Equated Delta*	*SD Equated Delta*	*Mean Biserial r*
1970	11.7	2.9	.47	11.9	2.8	.46	11.9	2.8	.45
1971	11.7	2.9	.47	11.8	2.9	.46	11.8	2.9	.45
1972	11.7	2.9	.47	11.5	2.9	.47	11.8	3.0	.44
1973	11.7	2.9	.47	11.7	3.1	.47	11.7	3.1	.44
1974	11.4	3.3	.48	11.5	3.3	.46	11.5	3.2	.46
1975	11.4	3.3	.48	11.4	3.4	.48	11.4	3.2	.48
1976	11.4	3.3	.48	11.3	3.4	.49	11.4	3.1	.46
1977	11.4	3.3	.48	11.3	3.3	.46	11.3	3.1	.46
1978	11.4	3.3	.48	11.4	3.1	.47	11.4	3.4	.46
1979	11.4	3.3	.48	11.4	3.1	.46	11.5	3.3	.44
1980	11.4	3.3	.48	11.1	3.3	.49	11.3	3.2	.46
1981	11.4	3.3	.48	11.2	3.2	.50	11.3	3.4	.48
1982	11.4	3.0	.46–.50	11.5	3.0	.51	11.3	2.9	.45
1983	11.4	3.0	.46–.50	11.4	3.0	.50	11.2	2.8	.48
1984	11.4	3.0	.46–.50	11.3	3.1	.49	11.4	2.9	.48

[a]Specified in terms of final-form items, which are included in the total-score criterion.

Table 5. Specified and Actual Item Statistics for November and December SAT-Mathematical Forms from 1970 to 1984

	Specified			*November Actual*			*December Actual*		
Year	*Mean Equated Delta*	*SD Equated Delta*	*Mean Biserial r*[a]	*Mean Equated Delta*	*SD Equated Delta*	*Mean Biserial r*	*Mean Equated Delta*	*SD Equated Delta*	*Mean Biserial r*
1970	12.5	3.1	.53	12.3	3.0	.52	12.3	3.0	.52
1971	12.5	3.1	.53	12.2	3.0	.54	12.3	3.0	.53
1972	12.5	3.1	.53	12.4	3.1	.56	12.3	3.1	.54
1973	12.5	3.1	.53	12.6	3.5	.54	12.5	3.0	.53
1974	12.2–12.3	3.1–3.3	.53	12.5	3.6	.53	12.1	3.0	.52
1975	12.2–12.3	3.1–3.3	.53	11.8	3.3	.58	11.8	3.1	.54
1976	12.2–12.3	3.1–3.3	.53	12.1	3.2	.56	12.0	2.9	.56
1977	12.2–12.3	3.1–3.3	.53	12.2	3.2	.54	12.4	3.5	.55
1978	12.2–12.3	3.1–3.3	.53	12.1	3.1	.52	12.2	3.3	.52
1979	12.2–12.3	3.1–3.3	.53	12.3	3.3	.54	12.2	3.2	.52
1980	12.2–12.3	3.1–3.3	.53	12.2	3.1	.54	12.2	3.1	.53
1981	12.2–12.3	3.1–3.3	.53	12.1	3.5	.56	12.4	2.9	.53
1982	12.2–12.3	3.1–3.3	.53	12.3	3.3	.55	12.1	3.1	.55
1983	12.2–12.3	3.1–3.3	.53	12.1	3.4	.55	12.0	3.2	.55
1984	12.2–12.3	3.1–3.3	.53	12.6	3.3	.54	12.2	3.5	.57

[a]Specified in terms of final-form items, which are included in the total-score criterion.

The Pre-October 1974 SAT

In order to clarify the nature of changes made to accommodate the TSWE in October 1974, the following general description of the SAT is provided for the period immediately preceding, i.e., January 1961 to September 1974. Item types are listed in their order of occurrence in the test sections. The appendix contains item-type examples.

The pre-October 1974 SAT included 150 items administered in 150 minutes. In addition, a 30-minute variable section contained equating or pretest questions that were not counted in the test takers' scores. From the test taker's perspective, the test lasted three hours.

The verbal portion of the SAT, containing 90 items, was administered in two separately timed sections. The first section, containing 50 items, required 45 minutes. There were 25 reading comprehension items based on five passages, plus 8 sentence completion items, 8 antonyms, and 9 analogies. The second SAT-verbal section contained 40 items and required 30 minutes. Ten items were allocated to each of the four verbal item types. The reading comprehension items were based on two passages.

The complete verbal portion of the SAT included seven reading passages, one each of the following types: narrative, biological science, physical science, argumentation, humanities, synthesis, and social studies.

The mathematical portion of the SAT, containing 60 items and administered in 75 minutes, was also divided into two separately timed sections. The first section had 25 regular math items requiring 30 minutes. In the second section there were 17 regular math items plus 18 data sufficiency items requiring 45 minutes.

The complete SAT was packaged in booklets arranged in one of two ways. For arrangement A, the sequence was verbal section 2, verbal section 1, the variable section, mathematical section 2, and mathematical section 1. The sequence in arrangement B was verbal section 2, verbal section 1, mathematical section 1, mathematical section 2, and the variable section. Only one arrangement was used at each administration of the test.

MODIFICATIONS TO THE SAT AFTER TSWE

The TSWE is a 30-minute test. Its introduction in October 1974 required shortening the SAT by an equal amount of time. To accomplish the reduction, each portion of the SAT was shortened 15 minutes (see Table 6 for a summary of changes).

The new SAT-verbal sections contained 85 items administered in 60 minutes. The first section, 30 minutes long, contained 45 items: 10 reading comprehension items based on two passages, 10 sentence completion items, 10 analogy items, and 15 antonyms. The second section, also 30 minutes long, contained 40 items: 15 reading comprehension items based on three passages, 5 sentence completion items, 10 analogies, and 10 antonyms. Two new SAT-verbal subscores were reported: a reading score based on reading comprehension and sentence completion items, and a vocabulary score derived from antonym and analogy items.

Other details regarding the modified verbal test are as follows:

- Item difficulty specifications were reduced from a mean delta of 11.7 to 11.4 in order to decrease the average difficulty of the test.
- The previous unimodal distribution of item difficulties was made bimodal to maintain discrimination at the upper end of the scale. As a result, the specified standard deviation of deltas was increased from 2.9 to 3.3. The number of difficult items (delta 15 or above) was increased from 14 out of 90 to 16 out of 85. Correspondingly, easy items (delta 8 or below) were increased from 18 out of 90 to 25 out of 85.
- The mean biserial (item total score) correlation was increased from .42 to .43 based on pretest statistics.
- The maximum number of words in reading comprehension passages for the test as a whole was reduced from 3,500 to 2,000–2,250.
- The synthesis passage dealing with the relationship between science and humanities was deleted.

Table 6. Summary of Changes Made to SAT Item Types, Content, and Test Format from March 1970 to January 1985

Beginning Date	*Change*
March 1973	One minority-relevant reading passage included in at least one SAT-V form administered during the testing year
October 1974	Two 30-minute SAT-V sections (45 and 40 items, respectively) introduced in place of one 45-minute section (50 items) and one 30-minute section (40 items) Two 30-minute SAT-M sections (25 and 35 items, respectively) introduced in place of one 30-minute section (25 items) and one 45-minute section (35 items) 30-minute Test of Standard Written English (50 items) introduced and administered in test booklet with the SAT Number of SAT-V item types changed: ○ Number of antonyms increased from 18 to 25 ○ Number of analogies increased from 19 to 20 ○ Number of sentence completions reduced from 18 to 15 ○ Number of reading comprehension passages reduced from 7 to 5; number of reading comprehension items reduced from 35 to 25 Length and content of reading passages altered: ○ Total words in reading passages reduced from a maximum of 3,500 to 2,000–2,250 ○ Deletion of syntheses and one of two science passages (biological or physical science at the discretion of test assembler) Reading (based on reading comprehension and sentence completion items) and vocabulary (based on antonym and analogy items) subscores introduced: ○ Reading and vocabulary items required to have similar mean item difficulties and standard deviations of item difficulties ○ Difficult vocabulary not used in sentence completion items ○ Number of more difficult sentence completion items increased Number of SAT-M item types changed: ○ Number of regular mathematics items reduced from 42 to 40 ○ 20 quantitative comparison items added ○ 18 data sufficiency items deleted Six (rather than one of two) fixed section orders used at each test administration
November 1975	To attempt to reduce speededness: ○ 10 reading comprehension items (based on two passages) moved from end to middle of SAT-V 45-item section, and 15 reading comprehension items (based on three passages) moved from the middle to the end of SAT-V 40-item section ○ 20 quantitative comparison items moved to middle of SAT-M 35-item section
1977–78	Slight reduction in the number of SAT-M items requiring a more complex knowledge of geometry Virtual elimination of the generic "he" from the SAT-V
December 1977	One minority-relevant reading passage included in each new form of the SAT-V

Table 6. (Continued)

Beginning Date	*Change*
October 1978	Number of reading passages increased from five to six: ○ Three 200–250-word passages replaced two 400–450-word passages ○ Second science passage returned to test ○ Two to four rather than five items used for each shorter passage One of two fixed section orders used at each administration
1979–80	Seven rather than five new forms produced each year to fulfill the requirements of test disclosure
1980	Test sensitivity guidelines implemented; tests reviewed to eliminate any material offensive and patronizing to women and minority groups; representation in test items of contributions of women and minority groups to American society; improvement in the ratio of male-female references
October 1980	One of three fixed section orders used at each administration
1981–82	Nine or ten new forms produced each year to fulfull the requirements of test disclosure

- Only one science passage was used. At the discretion of the test assembler, it could be either a biological or a physical science passage.
- Along with the introduction of two subscores, reading and vocabulary, items were required to have mean item difficulties similar to that of the whole test.
- Difficult vocabulary was not used in the sentence completion items.
- The number of more difficult sentence completion items was increased.

The new mathematical test contained 60 items administered in 60 minutes. The first section, 30 minutes long, had 25 regular math items. In the second section, there were 15 regular math items and 20 four-choice quantitative comparison items. The latter item type replaced 18 five-choice data sufficiency items used in previous editions of the test. Data sufficiency items took more time to answer.

Other details regarding the modified mathematical test are as follows:

- SAT-mathematical difficulty specifications were reduced from a mean delta of 12.5 to 12.2.
- The standard deviation of item difficulties was increased from delta 3.1 to 3.2.
- The number of difficult items (delta 15 or above) remained at 15 and the number of easy items (delta 8 or below) increased from 7 to 9.
- The mean biserial correlation stayed at .47 based on pretest statistics.
- There was a decrease in the number of items specified for ability level five due to the elimination of data sufficiency items, all of which were at this level.
- There have been one or two fewer geometry items in SAT-mathematical sections since 1974.

The TSWE did not become a permanent feature of the ATP until 1977. However, several modifications in the SAT made between 1974 and 1977, together with changes made in 1978 and 1980, can be regarded as further fine-tuning of an SAT that would operate in conjunction with the TSWE. For example, regarding SAT-verbal sections, adjustments were made pertaining to reading comprehension items in November 1975. In the first verbal section, 10 reading comprehension items based on two passages were relocated from the end to the middle of the section between two blocks of sentence completion items (five in each block). In the second verbal section, to reduce speededness and to improve item spacing in the test book, 15 reading comprehension items were moved from the middle to the end of the section. Later, in October 1978, the number of reading passages in the second verbal section was increased from three to four by using 220–250-word passages instead of 400–500-word passages. And once again the test contained both a biological and a physical science reading passage. The use of two kinds of science reading passages was maintained through 1985.

In January 1982 the statistical specifications for SAT-verbal sections were changed. The number of difficult items (delta 15 or above) was reduced from 16 to 8, the number of moderately difficult items (delta 13–14) was increased from 16 to 24, and the number of items below delta 8 was reduced from 15 to 13. These changes were made to strengthen the test's measurement power at the middle-to-upper parts of the score range.

Beginning in November 1975 and continuing to the present, a modified arrangement of items in the second section of the mathematical test has been used. Fifteen regular math items were split into one group of seven and a second group of eight, which were then presented before and after the 20 quantitative comparison items. Recall that the earlier pattern was 15 regular math items followed by 20 quantitative comparison items. The new arrangement allowed the easier and less time-consuming quantitative comparison items to be reached sooner because the more difficult and slower regular math items were placed at the end of the section.

As a new instrument incorporated within the structure of the SAT, the TSWE changed how the SAT was packaged for test administration purposes. Beginning in October 1974 the SAT was packaged in booklets having six variations of the following sequence of sections: verbal 1, mathematical 1, TSWE, verbal 2, mathematical 2, and variable. The first booklet followed the sequence as shown. The second booklet began with mathematical 1 and continued in sequence so as to end with verbal 1, and so on for the six variations. The procedure was known as scrambling. Moreover, the booklets were collated (spiraled) for shipment to test centers. The spiraling procedure meant that at the test center the first student received the first variation, the second student the

second variation, and so on. The procedure was intended to reduce the possibility of copying at the test administration.

This method of scrambling the six sections ended in October 1978 because of its complexity and the belief that certain scrambles were less desirable than others. Thus from October 1978 through June 1980 the test was packaged in booklets containing one of the following sequences: verbal 2, mathematical 2, variable, verbal 1, mathematical 1, and TSWE; or verbal 1, mathematical 1, TSWE, verbal 2, mathematical 2, and variable.

In October 1980 the first sequence was dropped and the second sequence was used with two others: verbal 2, variable, mathematical 1, TSWE, mathematical 2, and verbal 1; or verbal 1, mathematical 1, verbal 2, variable, mathematical 2, and TSWE. Note that mathematical 2, the 35-item section with 20 four-choice quantitative comparison items, was now always located in the fifth position. This arrangement allowed a standard answer sheet with four and five response options to be used at all administrations. One of these arrangements prevailed at each administration from October 1980 through 1985.

OBSERVATIONS PERTINENT TO THE INTRODUCTION OF TSWE

Following the reporting of scores for each administration in which a new form of the SAT is used, ETS prepares a test analysis report. The data source is the answer sheets for a sample of approximately 2,000 test takers. Before 1981, the samples were drawn to statistically represent the total population of test takers. Since 1981, the samples have been restricted to high school juniors and seniors and have excluded junior high, high school sophomore, and adult test takers. Test analysis data permit observations to be made regarding the statistical characteristics of the test in response to the changes discussed in the previous section. The test analysis data referred to here came primarily from the reports of the November and December test forms, the two forms taken by the most graduating seniors.

In addition to the routinely prepared systematic test analyses, special studies of particular issues are conducted by ETS from time to time. These studies have also been drawn upon for additional observations regarding the major modifications in the SAT during the 1970–85 era.

Test Difficulty Specifications

For the most part, November and December SAT-verbal and SAT-mathematical forms met specifications from 1970 through 1984 (see Tables 4, 5, and 7). Mean deltas for SAT-verbal forms were, indeed, reduced from the level of 11.7, which had been specified before October 1974. November and December SAT-verbal forms before 1974 were likely to be more difficult than specified and more likely to be easier

Table 7. Summary of Statistical Characteristics of November and December SAT Forms from 1970 to 1984

Index	*Observation*
Mean Equated Delta	Drop in test difficulty noticeable in 1974 for both the SAT-V and the SAT-M
	Dec. 1980–84 SAT-V forms easier than previous forms
	SAT-V forms harder than specified before 1974 and easier than specified from 1974 on
	SAT-M forms easier than specified before 1974; sometimes easier and sometimes harder than specified from 1974 on
Distribution of Equated Deltas	Few large deviations from intended distributions
	No systematic trends in SDs of equated deltas; 6 of 30 SAT-V delta SDs and 9 of 30 SAT-M delta SDs were out of range
Mean Biserial r	Nov. values higher than Dec. values for both the SAT-V and the SAT-M
	SAT-V values close to specified values
	Nov. SAT-V mean correlations from 1980 to 1984 higher than specified; Dec. SAT-V correlations tended to be lower than specified from 1970 to 1984
	SAT-M values fluctuated more than SAT-V values
	SAT-M values tended to be higher than specified from 1970 to 1984; values from Nov. 1975 and Dec. 1984 forms especially high (by .04 to .05)
Score Conversions	Values somewhat inconsistent with mean equated deltas
	Drop in test difficulty noticeable in 1974 for both the SAT-V and the SAT-M
	Most SAT-V forms from 1980 to 1984 easier than previous forms
	Most SAT-M forms from 1981 to 1984 easier than previous forms
	Relatively less form-to-form variation in score conversions in 1974–78 for the SAT-V and in 1974–81 for the SAT-M
Mean Adjusted Proportion Correct (and Mean Observed Delta)	Consistent patterns evident for mean adjusted proportions correct and mean observed deltas
	SAT-V and SAT-M relatively difficult for test takers throughout period; Nov. mean proportions correct ranged from .40 to .44; Dec. from .36 to .39
	Increase in Nov. SAT-V mean proportion correct in 1974, reflecting easier test
	Increase of .01 in Nov. SAT-V mean proportion correct from 1975–79 to 1980–84
	Decrease in Dec. SAT-V mean proportions correct to low of .36 in 1977–79 and increase from 1982 to 1984 to high of .39

Table 7. (Continued)

Index	*Observation*
	Decrease in test difficulty in 1974 not reflected in SAT-M mean proportion correct
	Slight decrease in Nov. SAT-M mean proportions correct from 1970–73 to 1974–80 and increase in means from 1980 to 1984
	Decrease in Dec. SAT-M means from 1971 to 1977 and increase from 1981 to 1984
Percentage Completing 75% of Test (and Ratio of Not Reached to Section Score Variance)	Reasonably consistent patterns between percentage of test takers completing 75% of test and ratio of variances
	SAT-V and SAT-M more speeded for Dec. test takers than for Nov. test takers
	Longer verbal 1 section more speeded than verbal 2 section in Dec. except for Dec. 1974 and 1984 forms
	Verbal 1 tended gradually to become more speeded for Nov. test takers from 1975 on and for Dec. test takers from 1974 to 1983
	Considerable fluctuation in verbal 2 values from 1974 on, suggesting form-specific speed factors
	Verbal 2 more speeded than verbal 1 in 1974 with introduction of shortened SAT
	Verbal 2 relatively unspeeded except in 1974
	Shorter mathematical 1 section unspeeded from 1974 on relative to the mathematical 1 sections in 1970–73 forms
	Mathematical 2 section relatively unspeeded except in Nov. 1974 and December 1972–75; format change introduced in 1975 apparently helped reduce speededness
Internal Consistency Reliability	Reliabilities at or above .91 for both SAT-V and SAT-M from 1970 to 1984
	Decrease in reliability in 1974 for all but Dec. SAT-V form, but reliabilities higher in succeeding years
	1980–84 Nov. SAT-V reliabilities higher by .01; SAT-M reliabilities from 1978–84 increased or remained stable
	High SAT-M reliabilities in Nov. 1975, Nov. and Dec. 1976, and Dec. 1984; low reliability in Nov. 1978
Adjusted Internal Consistency Reliability (SD = 100)	Adjusted reliabilities lower than unadjusted because SD of 100 was lower than actually observed
	Adjusted reliabilities less variable than unadjusted reliabilities
	Decreases in reliability in 1974 for all but Dec. SAT-V form; no pattern of depressed reliabilities later, however
	Relatively stable SAT-V reliabilities except for Dec. 1973 form—most reliabilities ranged from .90 to .91

Table 7. (Continued)

Index	*Observation*
	Slight upward trend in Dec. SAT-V reliabilities from 1970 to 1984, but downward trend from 1981 to 1984
	Downward trend in Nov. SAT-V reliabilities from 1982 to 1984
	Adjusted reliabilities more variable for the SAT-M than the SAT-V, especially for Dec. forms
	No systematic patterns evident in SAT-M reliabilities—most reliabilities ranged from .88 to .89
	Dec. SAT-M reliabilities .01 higher than Nov. reliabilities in 1982, 1983, and 1984
	Dec. 1976 reliability stood out as too high
Test-Retest Correlation	High correlations from 1970 to 1984, ranging from .87 to .89 for most values
	Relatively low SAT-V correlations observed for March/April to Nov. from 1974 to 1978 (.88) and for March/April to Dec. from 1973 to 1980 (.87)
	No obvious trends in SAT-M correlations—correlations in later years similar to those in earlier years
Correlation Between SAT-V and SAT-M	Considerable fluctuation in correlations—.62 to .71
	General downward trend in 1970–84 period from .68 to .66
	Nov. correlations relatively stable except for unusually large decrease then increase in Nov. 1979 and Nov. 1980
	Greater fluctuation in Dec. correlations
	Large decrease in Dec. 1974
Correlation Between SAT-V and TSWE	Most correlations between .78 and .79
	Nov. correlations slightly lower than Dec. correlations
	Nov. correlations increased and then leveled off
	Dec. correlations increased to high of .81 in 1978 and then decreased to previous levels
	Low point (.75) in Nov. 1974
Correlation Between SAT-M and TSWE	Most correlations between .62 and .64
	Relatively low correlations (.59) in both Nov. and Dec. 1974
	Increase in Nov. correlations from .59 to .63 from 1974 to 1976 reached a high of .64 in 1981
	Increase in Dec. correlations from .59 to .65 from 1974 to 1977
	Unusually low correlation of .55 in Dec. 1981
Correlation of Sections 1 and 2	Section correlations corrected for attenuation for both the SAT-V and the SAT-M ranged from .97 to 1.00
	SAT-M corrected correlations slightly higher than those for the SAT-V
	SAT-V correlations more variable; no pattern

BEST COPY AVAILABLE

Table 7. (Continued)

Index	*Observation*
	SAT-M sections slightly more homogeneous from 1974 on than before
Correlation of Reading and Vocabulary	Correlations corrected for attenuation for reading ranged from .92 to .96
	Most corrected correlations varied only slightly from .94
	Uncorrected correlations were variable and averaged .80

than specified from October 1974 on. Only in November 1980 was the difference greater than .2 between the actual and specified mean deltas.

Mean deltas for SAT-mathematical forms before 1974 were more likely to be easier than specified. Thereafter, the forms were sometimes easier and sometimes harder. The largest discrepancies were on the order of .4 in either direction. No systematic trends in delta standard deviations were observed for SAT-verbal or SAT-mathematical forms. The differences were relatively minor and in general within .2 of specifications.

Biserial Correlations

Generally the mean item-total test biserial correlations for November and December SAT-verbal forms came close to specified values. November correlations were higher than those for December and, indeed, in 1980 and later, they were higher than specified. December SAT-verbal biserial correlations tended to be lower than specified throughout the 15-year period. SAT-mathematical biserials fluctuated more than SAT-verbal biserials. The forms with the largest deviations were those for the November 1975 and December 1984 administrations, when biserials were .05 and .04 higher than specified. Although tests with high biserials tend to have higher reliabilities, they may not maintain the desired breadth of coverage.

Score Conversions

A given raw score on a test designed to be easier than a previous form should yield a lower scaled score than the earlier form. The converse is also the case, i.e., an increase in mean equated delta (higher difficulty) should yield higher scaled scores corresponding to given raw scores. However, delta equating and score equating are independent activities. Did they produce consistent results? Two indices are pertinent to this issue: (1) scaled scores corresponding to the midpoints of the raw score ranges, or scaled score ranges corresponding to selected raw scores; and (2) comparisons of mean equated deltas.

For SAT-verbal forms, both indices show reasonably consistent results for the planned decrease in difficulty in 1974 as well as the unplanned decrease in mean equated deltas in November 1972 and November 1980. Moreover, both indices show that from 1980 to 1984, many SAT-verbal forms were easier than previous forms. There were discrepancies. The two indices did not match in December 1977 and November 1982.

For SAT-mathematical forms, although both indices reflected the planned decrease in test difficulty in 1974, there were inconsistencies. Score conversion data for 1981 onward indicated that the test was relatively easier. However, this trend was not matched with mean equated delta scale data. The implication is that the delta scale may have drifted upward relative to the score scale, which for score data is more accurate.

When one examines the entire range of scores and not just scores at the midpoint, as was the case in the preceding discussion, one notes:

- When the shortened (easier) verbal test was introduced, scaled scores corresponding to raw scores decreased.
- In the upper score ranges, SAT-verbal scores drifted downward from 1974 to 1985.
- SAT-mathematical scaled score ranges drifted downward from January 1982 to 1985 despite unchanged statistical specifications.

Relative Test Difficulty

The abilities of test takers have not been stable over the years. Relative test difficulty refers to the difficulty of a test form relative to those who took that form. As the abilities of test takers increase, relative test difficulty should decrease, and vice versa. In the case of the SAT, the observed mean delta is one index of relative difficulty. A second such index is the mean raw score (for the SAT, the raw score is actually the number right minus a fraction of the number wrong) divided by the number of items in the test. This ratio is referred to as the mean adjusted proportion correct.

Given the decline in average SAT scores in the 1970s to the early 1980s and their gradual increase from 1982 to 1985, the test should correspondingly have been relatively more difficult in the earlier period and relatively easier in the later period. Actual trends in relative test difficulty occasionally patterned changes in average test-taker ability, but they also reflected changes in test difficulty specifications. More recent forms measured the ability of the average test taker as well as or better than earlier forms. Nonetheless, the SAT remained difficult for the average test taker.

Speededness

Four indices are used to determine speededness (i.e., the extent to which test takers are unable to complete a test section within the allotted time):

1. The percentage completing 75 percent of the section.

2. The percentage completing 100 percent of the section.
3. The ratio of items not reached to total test score variance.
4. The mean and standard deviation of the number of items not reached.

The first three indices take section length into account. The second index is problematic because one very difficult item at the end of the test can greatly reduce the percentage completing the test. With the fourth index section length can be taken into account by dividing by the number of items in the section. ETS practice is to regard a test as unspeeded if virtually all test takers complete 75 percent of the items in a timed section and 80 percent of the test takers complete all the items in a timed section.

The issue of speededness arose in connection with shortening the SAT in 1974, when the modifications to the test were intended to save time. Speededness was also an issue in 1978 when the number of reading passages was increased from five to six and their length and the number of items per passage simultaneously decreased. Recall, too, that the addition of a second science reading passage was included in the 1978 modifications. Investigations indicate that the shorter verbal sections and the longer mathematical sections tended to be relatively unspeeded for most students tested in November and December from 1974 to 1984. These sections became more speeded with the shortening of the test in 1974.

In 1975, however, the change in the order of item types within the SAT-verbal and SAT-mathematical sections reduced speededness to previous levels. The longer SAT-verbal section gradually became more speeded. The shorter SAT-mathematical section gradually became less speeded and was relatively unspeeded from 1976 on. The addition of a reading passage and the shortening of the reading passages in 1978 did not seem to make the SAT-verbal sections more speeded.

Reliability

Reliability concerns the extent to which tests measure true differences in the attribute being measured rather than variations due to chance or factors other than those being tested. Assessment of test reliability essentially involves a comparison of the true variance in the distribution of test scores and the variance in scores due to random errors in the testing process. Two kinds of reliability estimates of SAT scores are available. Internal consistency estimates assess the extent to which items in a test measure the same underlying factor. This estimate does not account for differences among test forms and thus does not include the effects of equating. Test-retest reliability, on the other hand, assesses the degree to which a second test administration yields similar scores for the same individuals. For the SAT, these estimates are based on correlations of scores on alternative forms of the test taken by high school juniors tested in the spring who take the test again in the fall of their senior year. This estimate is attenuated because of changes in the abilities of test takers that occur over time.

Reliability coefficients range from 0.00 to 1.00. As a general rule reliable tests have an internal consistency index of .90. Test-retest reliability indices, however, tend to be slightly lower.

- *Internal consistency reliability.* The Dressel (1940) adaptation of the Kuder-Richardson Formula 20 is used to calculate SAT internal consistency reliability. During the 15-year period from 1970 to 1984, SATs, both verbal and mathematical, maintained high levels of internal consistency as measured by coefficient alpha. Except for the November 1978 SAT-mathematical form, all reliability coefficients were at or above .90 for both SAT-verbal and SAT-mathematical forms. Although reliability levels dropped in 1974 following the shortening of the test, the reliabilities were higher in succeeding years, thus discounting any effects of shortening the test on reliability.
- *Test-retest reliability.* Correlations of junior year to senior year performance remained relatively stable between 1970 and 1984. For approximately 200 comparisons of both SAT-verbal and SAT-mathematical forms, the junior-senior correlations ranged from .87 to .89. There were five exceptions, all at the .86 level: one in 1979 for the verbal sections, three before 1972 for the mathematical sections, and another for the mathematical sections in 1980.

Correlational Patterns

Trends in the correlations of SAT-verbal and SAT-mathematical scores with other variables, with each other, and with SAT sections and subsections merit consideration. If stable over the years, such correlational patterns would provide indications that the test measured the same thing to the same degree of precision. These correlations, of course, can be affected by changes in the test-taking population.

- *Correlations among SAT-verbal, SAT-mathematical, and TSWE scores.* For November and December SAT forms from 1970 to 1984, the correlations of SAT-verbal and SAT-mathematical scores ranged from .62 to .71, indicating that the two sections were measuring different constructs. There was also a slight downward trend for both November and December test takers during the 1970–84 period, with the correlations dropping by .02 for both administrations. As expected, because both tests measure verbal attributes, correlations of SAT-verbal and TSWE scores were higher (i.e., high .70s) than those between SAT-mathematical and TSWE scores (low to mid .60s) or between SAT-verbal and SAT-mathematical scores.

- *Correlations of section 1 and 2 SAT-verbal and SAT-mathematical scores.* Correlations of section scores, corrected for unreliability, ranged from .97 to 1.00. SAT-mathematical correlations were slightly higher than SAT-verbal correlations. The data suggest that the SAT-mathematical sections became even more homogeneous after the changes in October 1974. That pattern was not apparent for SAT-verbal sections.
- *Correlations of reading and vocabulary subscores.* The correlations of the two verbal subscores, corrected for attenuation, ranged from .92 to .96, indicating that essentially the same underlying construct was being measured.

MODIFICATIONS TO ACCOMMODATE TEST DISCLOSURE LEGISLATION

In June 1979 New York State enacted test disclosure legislation effective January 1980. The legislation required the disclosure of all postsecondary admission tests after each administration. For a reasonable fee, test takers could request a copy of the questions, the correct answers, and a copy of their answer sheets. Explanations were required of what each test was designed to measure, its limitations, and how it was scored. Subsequent amendments exempted low-volume tests from annual disclosure but required disclosure of at least one non-Saturday administration. Disclosure began in 1980–81 in New York State with the release of four Saturday SATs and one SAT from a Sunday administration.

Before passage of the legislation, the contents of test forms could be regarded as secure. Test questions and test forms could be and were reused in more than a single administration. The legislation effectively eliminated the reuse of disclosed test forms and necessitated an increase in SAT development. The domino effect was substantial. Additional test development meant increased pretesting, larger staffing, increased use of external personnel to write items, and production of training materials for outside item writers. Procedures were needed for the Question-and-Answer Service. More thorough reviews of tests before their operational use became necessary, and so on. The following changes were made in policies and procedures pertaining to the SAT in order to meet the requirements of disclosure legislation:

- *Final SAT forms development increased.* During the period 1970–78 approximately five new SAT forms were produced annually. Development increased to seven new forms a year in 1979–80. From 1981 to 1985 nine to ten new forms were produced annually.
- *Pretesting increased.* In 1979–80 approximately 40 verbal and mathematical pretests were produced each year. Production increased to 75 verbal and mathematical pretests in 1980–81, and jumped again in 1981–82 to 100 verbal and mathematical pretests.
- *Test development staffing increased.* Between 1970 and 1977 SAT-verbal test developers increased from three to five. In 1978–79 the staff increased to eight, followed by an increase to ten in 1980–81 and to fourteen in 1984–85. During 1970–79 final forms assemblers increased to five a year. There were similar increases in SAT-mathematical test development staff during the same periods.
- *External item writing increased.* Before 1980 SAT-mathematical item writing was done by in-house ETS staff or former staff. Since 1980 outside item writers have been used, necessitating the preparation of training materials, training exercises, and screening of personnel before awarding work assignments. Outside SAT-verbal item writing increased less than that for SAT-mathematical sections. Reading comprehension sets had routinely been obtained from outside writers even before the disclosure requirements. However, in order to maintain better control of overlap with materials previously written and disclosed, antonyms, analogies, and sentence completion item writing continued to be done primarily by ETS staff.
- *External professional oversight and review of the SAT changed.* In the 1970–85 era there were some modifications in the arrangements for external professional advice regarding the SAT. In the early years, 1970–72, there was a College Board–appointed Committee of Examiners in Aptitude Testing that consisted of professional educational measurement specialists—college and university faculty members—whose advice was mainly oriented toward psychometric issues. Meanwhile ETS had itself acquired a substantial professional staff whose expertise overlapped that of the College Board committee. Therefore the Committee of Examiners in Aptitude Testing was disbanded in 1972.

There was no external review group until 1977, when the College Board appointed a committee of college and high school educators and administrators and charged it with providing advice regarding SAT policy and program issues. Three members of the committee also reviewed each new edition of the SAT by mail. Thus came into being the first regularly scheduled external reviews of the test. The College Board's Scholastic Aptitude Test Committee continued through 1985.

In the early 1980s a flawed test item, publicized because of test disclosure, led to an expansion of the external review process, which was formalized in 1984. A panel of 15 subject-matter experts was appointed to review new forms of SAT-mathematical sections beginning with the October 1984 administration. Five of these content experts reviewed each new form of the mathematical test, and three of the five then met with ETS test assemblers. After the test was revised on the basis of the external reviews, it was reviewed by three ETS test development staff plus the test assembler. The same procedure was initiated in October 1985 for new SAT-verbal forms using a different 15-member external review panel.

As a consequence of the changes described above, from 1970 to 1985, internal reviews per final form increased

from five to nine, and the total number of external reviewers went from zero to eight.

• *Question-and-Answer Service initiated.* New York disclosure legislation required that test takers receive, upon request and for a reasonable fee, a copy of the test questions, the correct answers, a copy of their answers, and information to allow raw score to College Board score conversions. In the spring of 1981, the College Board initiated its Question-and-Answer Service for a fee of $6.50. Although instituted in response to the legislation, since 1982–83 the service has been available to all test takers for at least one major administration of the SAT each year. About 20,000 test takers (less than 2 percent of those tested) used the service in 1982–83. In 1984–85 there were 16,051 participants in the service, representing about 1.3 percent of the test takers.

• *Test information for students furnished.* Students who register to take the SAT are furnished with an information bulletin. In the 1970–85 era there were eight variations in the kinds and amount of information provided to students about the SAT. These variations were only coincidentally related to changes in the test itself, except, of course, when TSWE appeared on the scene in 1974–75, and in 1979–80 in anticipation of test disclosure requirements, but a thorough history seems appropriate.

In 1970–71, a 55-page bulletin was used. It contained 16 verbal items and 17 mathematical items with explanations. In addition there were 57 verbal and 36 mathematical items that were not explained. The 1971–72 bulletin changed only to the extent that one fewer explained mathematical item was included. In 1972–73 and 1973–74, to ensure that information about the test was received by all registrants, bulletins were mailed to them with their tickets of admission. The bulletin was cut to 15 pages to hold down mailing expenses. Ten verbal and seven mathematical items were provided, all with explanations.

In 1974–75 and 1975–76 the bulletin had 12 pages. Explained verbal and mathematical items numbered 4 and 8, respectively. In addition 21 verbal and 8 mathematical unexplained items were included. The allocation of explained items changed in 1976–77: 8 verbal and only 7 mathematical items were included, along with 15 verbal and 14 mathematical unexplained items. Although the 16-page bulletin for 1977–78 contained no explained items, it had 30 each unexplained verbal and mathematical test items.

There was an expansion of explanatory material beginning in 1978–79. Additionally in that year and continuing through 1981–82 a full-length unexplained verbal and mathematical test was published. The bulletin jumped to 48 pages and included 21 verbal and 17 mathematical explained items. A mathematical review section, indicating the skills and content emphasis of the test, was first used in the bulletin for 1978. That kind of material was used continuously through 1985.

Beginning in 1982–83 and continuing through 1984–85 comprehensive test information was available in a 62-page publication titled *Taking the SAT: A Guide to the Scholastic Aptitude Test and Test of Standard Written English.* It contained 23 explained verbal and 20 explained mathematical items. At the beginning of each academic year this free publication was shipped routinely to secondary schools for students who planned to take the SAT. More recently the College Board has offered a paid publication, *10 SATs,* which, as the title indicates, has actual SATs complete with answer keys, raw score to scaled score conversions, and advice about how to prepare for the test.

OBSERVATIONS REGARDING TEST DISCLOSURE

Postadministration security of test forms, a traditional College Board policy before 1980, allowed test items and test forms to be reused in subsequent administrations. Disclosure legislation forced a change in this policy. As previously indicated, there were consequences pertaining to the amount of new test development as well as the staffing required to produce the added volume of test material.

It also seems clear that disclosure has increased openness regarding the nature of the SAT. Beyond that, it has contributed to test-taker and general public knowledge of how, in general, the test is scored and reported. In addition, how each question on a particular test form is scored is now made public. Although the kinds and amounts of informational materials provided to test takers varied during the 1970–85 era, after disclosure legislation the College Board complied with the legal requirements, and more, in releasing information about the SAT.

It is also likely that disclosure legislation has been responsible for increased attention being given to preadministration review of the SAT by external experts and ETS test development staff. If this is true, and flawed test items are now even more rare than previously, disclosure has been a plus.

On the negative side, disclosure probably has had a subtle impact on the quality of some items in the test. Test developers, now sensitive to working in a disclosure environment, may be less likely to use questions that call for fine distinctions, and items with very close distractors may not find their way into final test forms. The matter is not critical, but it is a difference.

MODIFICATIONS RELATED TO TEST SENSITIVITY REVIEW

The Rationale

Early in the 1970–85 era, an emerging concern at ETS was the need to ensure that all programs for which it provided testing services were tuned to changing societal values and attitudes. This concern extended beyond the measurement

characteristics of tests to their psychological attributes. More specifically, materials in tests should recognize "the varied contributions that minority members have made to our society" and there should be "no inappropriate or offensive material in the tests" (Educational Testing Service 1987, 4).

Efforts in the early 1970s were somewhat informal and exploratory. However, by 1980 ETS had formally adopted, as corporate policy, the ETS Test Sensitivity Review Process containing specific guidelines applicable to the development of tests (see Educational Testing Service 1986, 1987). For example, test developers were instructed to include in test specifications "requirements for material reflecting the cultural background and contributions of major population sub-groups." Another guideline required that test items, the test as a whole, as well as its descriptive materials not include language regarded as sexist, racist, or otherwise potentially offensive, inappropriate, or negative toward major subgroups. Moreover, the guidelines called for special consideration to be given to the perspectives of Asian/Pacific Island Americans, Black Americans, Hispanic Americans, individuals with disabilities, Native Americans/American Indians, and women. The guidelines can be extended to the elderly and members of other groups not mentioned specifically.

The Critical Elements

Several factors provide direction for the sensitivity review process:

- *Cultural diversity.* Tests and related publications should reflect the diversity of the test-taking population by recognizing the contributions of all groups to our culture and particularly by recognizing the contributions of women and minorities in all fields of endeavor.
- *Diversity of backgrounds among test takers.* Some test questions may touch emotional trigger points for some test takers but not for others. Sensitivity review should ensure that test materials dealing with disabilities, gender, or ethnicity, including incorrect answer choices, are developed with care.
- *Force of language.* Changing societal attitudes bring about differences in the acceptance of and response to words that are used in tests and supporting informational materials.
- *Changing roles.* Tests should recognize significant changes in our society, such as those of family patterns, and provide for a balanced portrayal of the roles and contributions of women and members of minority groups to ever-widening fields of endeavor.

The Process

Sensitivity review involves three steps. First, a preliminary review is requested by the test developer for screening test materials for sensitivity-related issues. Second, a mandatory final review is conducted during the regular editorial process after the test has been assembled. At this point the sensitivity reviewer notifies the test assembler in writing of any sensitive issues that the test has raised. Usually the two can resolve any issues raised by the reviewer. Third, arbitration by three staff members, who have no involvement with the test, occurs if the differences cannot be resolved by the reviewer and the test assembler.

The sensitivity review process is undergirded by the following written guidelines and procedures:

1. Evaluation guidelines—specific policies to assist all reviewers and editors in ensuring the fair treatment of all people in tests and publications and in applying the same standards to all programs and clients.
2. Evaluation requirements—detailed statements describing 11 perspectives that must be brought to bear on every test undergoing sensitivity review. For example, all group reference test items are reviewed from both their cognitive dimensions (the accuracy of the information in the items) and their affective dimensions (the positive or negative feelings the items may produce among groups taking the test).
3. More extended guidance in the form of numerous examples pertaining to unacceptable stereotypes, caution words and phrases, and special review criteria for women's concerns and for references to people with disabilities.

The adoption of test sensitivity review as ETS corporate policy meant that test development procedures for the SAT were modified to ensure conformity to the requirements of the policy. Indications of the impact of the policy on the SAT is presented in the discussion that follows.

OBSERVATIONS REGARDING TEST SENSITIVITY REVIEW

Minority-Relevant Reading Passages

Sensitivity concerns had early support by the College Board, particularly as related to the SAT. For example, pretesting of SAT-verbal reading passages relevant to minority groups began in 1970. A reading passage relevant to a minority group was first included operationally in the SAT in March 1973. From March 1973 until 1976–77, one such passage was included in at least one new form of the test each year. As of December 1977, a minority-group-relevant reading passage has been part of every new form of the test.

Gender References

The balance of gender references in the SAT-verbal sections becomes an issue primarily in sentence completion and

reading comprehension items. Cruise and Kimmel (1990) documented trends in content and gender references in SAT-verbal forms from 1961 to 1987. They found that the occurrence of sex-linked words in antonym items was so rare as not to merit inclusion in their study. Similarly, 90 percent of analogy items were noted to be gender neutral or to contain no human references, up from 88 percent at the beginning of their study.

They also found that since 1977–78 the use of the generic "he" has been virtually eliminated in sentence completion items and in reading comprehension passages. Comparing sentence completion items for years before and after test sensitivity review, they noted an increased representation of women to the point of near parity with men, together with an increased proportion of items that include humans.

Making the same time-period comparisons for reading comprehension questions, Cruise and Kimmel found recent indications of some decline in male references, indications of an increase in questions with no gender references, and a slight increase (from 1 percent to 4 percent) in female references. It is also of interest to observe that reading comprehension passages containing gender references declined from 76 percent in 1961–67 to 58 percent in 1982–87.

Cruise and Kimmel concluded with the following cogent observation:

> Although the ratio of male to female references has been reduced in recent years, the analysis of gender references indicates that throughout the period under study, the test has had a preponderance of male-oriented language and references. Much of this reflects the topics and activities that have entered into the language and into published writing that serves as the source of passages used in testing reading comprehension. There is no obvious criterion for judging whether the observed proportions of female and male references in the test are appropriate. Although there may be important social reasons for seeking a more balanced selection of language and reading passages, it is not clear whether the imbalance affects performance levels. (Cruise and Kimmel 1990, 12)

USE OF ITEM-RESPONSE THEORY EQUATING

As was noted earlier, despite being assembled to rigorous content and statistical specifications, SAT forms vary somewhat in statistical characteristics. Ensuring fairness to candidates taking different forms of the test and maintaining the recognized and accepted meaning of reported scores mandate the use of some procedure for making scores comparable. That procedure is called score equating. The process depends on linking each new form of the test to one or more previous forms using an anchor test. Generally speaking, the anchor test can be regarded as a miniature version of the SAT, except that the mathematical anchor test has only regular mathematics items. Since the anchor test (hereafter, also referred to as an equating section) is taken by current test takers and an earlier group whose scores have been placed on the College Board scale, the equating section provides the data for linking scores of the two test forms.

The data linkage is accomplished in the operational administration of the SAT by means of the format used in the test booklets. An SAT booklet has six sections: four operational SAT sections (two verbal and two mathematical) containing questions used to determine a test taker's raw score, the TSWE, and one of a number of variable sections that do not count toward the individual's score. These variable sections may be verbal or mathematical pretests or verbal or mathematical anchor tests linking the current form with previously administered forms or a future form.

The equating sections, linked to previously administered test forms, provide data for two independent estimates (which are usually averaged) of raw-to-scale score conversions of the new form. During the 1970s, there were some variations from the design just described (see Donlon 1984 for details).

Linear Equating

Before 1982 linear equating methods were used, namely the Tucker observed score, the Levine equally reliable, and the Levine unequally reliable equating methods. The last was used when total tests were of different lengths or when examinees in the two equating samples differed in ability. The assumption of linear methods is that the relationship between raw scores representing the same ability level on the tests to be equated can be graphically represented by a straight line. Using these methods, scores on two forms of the same test can be considered comparable if they correspond to the same number of standard deviations above and below the mean of the reference group of test takers. Performance on the anchor test is used to estimate the test takers' scores on one form or the other of a complete SAT.

Most equating during the 1970–85 era was done by the Tucker method. Levine methods alone were used for the SAT-verbal equating in December 1975. Levine methods were also used in either the first or second equating of the SAT-verbal form in November 1979 and for the December administrations of 1972, 1973, 1977, 1978, 1981, and 1983. Regarding the SAT-mathematical scores, Levine methods were used in either the first or second equating in December 1973, 1975, 1977, 1978, and 1981. Levine methods were used whenever the new-old form samples differed considerably in ability level. Table 8 provides a detailed chronology of the equating methods used for November and December SAT-verbal and SAT-mathematical equatings from 1970 to 1984.

Linear methods were used to equate SAT scores before the test was shortened, i.e., during the 1970–74 period. These methods were used primarily because the test was assembled to specifications that were the same during the period, thus assuring essentially parallel forms. Moreover, neither a fully satisfactory curvilinear equating method nor computer capability was available to handle curvilinear

Table 8. Equating Methods[a] Used for November and December SAT-Verbal and SAT-Mathematical Equatings from 1970 to 1984

	SAT-Verbal		*SAT-Mathematical*	
Year	*First Equating*	*Second Equating*	*First Equating*	*Second Equating*
		November Administrations		
1970	Tucker	Tucker	Tucker	Tucker
1971	Tucker	Tucker	Tucker	Tucker
1972	Tucker	Tucker	Tucker	Tucker
1973	Tucker	Tucker	Tucker	Tucker
1974	Tucker	Tucker	Tucker	Tucker
1975	Tucker	Tucker	Tucker	Tucker
1976	Tucker	Tucker	Tucker	Tucker
1977	Tucker	Tucker	Tucker	Tucker
1978	Tucker	Tucker	Tucker	Tucker
1979	Tucker	Levine	Tucker	Tucker
1980	Tucker	Tucker	Tucker	Tucker
1981	Tucker	Tucker	Tucker	Tucker
1982	IRT	IRT	IRT	IRT
1983	IRT	IRT	IRT	IRT
1984	IRT	IRT	IRT/Tucker[b]	Tucker/Tucker[b]
		December Administrations		
1970	Tucker	Tucker	Tucker	Tucker
1971	Tucker	Tucker	Tucker	Tucker
1972	Levine	Tucker	Tucker	Tucker
1973	Tucker	Levine	Tucker	Levine
1974	Tucker	Tucker	Tucker	Tucker
1975	Levine	Levine	Levine	Tucker
1976	Tucker	Tucker	Tucker	Tucker
1977	Tucker	Levine	Tucker	Levine
1978	Levine	Tucker	Levine	Tucker
1979	Tucker	Tucker	Tucker	Tucker
1980	Tucker	Tucker	Tucker	Tucker
1981	Tucker	Levine	Tucker	Levine
1982	IRT	IRT	IRT	IRT
1983	IRT	Levine	IRT	IRT
1984	IRT/IRT[b]	Tucker/Tucker[b]	IRT/IRT[b]	Tucker/Tucker[b]

[a]IRT refers to item-response theory.

[b]This equating went back to the two parent forms for the old form rather than to the old form itself. One of the parent forms for the old form used in this equating was the same as one of the parent forms used in the other equating. Therefore, in averaging the two equating lines, the equating to the common parent form was weighted half as much as the equating to the distinct parent form.

equating. Although equipercentile equating through an anchor test was performed in addition to linear equating, it was used only to check the curvilinearity of equating and as a basis for empirical "doglegs," which were linear line segments covering a small part, usually the upper end, of the score range.

When the SAT was shortened and the statistical specifications changed in 1974, linear methods continued to be used. The specified distribution of item difficulties for the SAT-verbal sections became more like that for the pre-1966 period except for an increase in the standard deviation of item difficulties. The bimodal distribution included more items at delta levels greater than or equal to 15 to ensure good measurement at the upper end of the scale despite the decrease in item difficulty. The distribution of SAT-mathematical difficulties shifted downward slightly to make the test less difficult, but otherwise it looked very much like the distribution for the earlier period.

IRT Equating

Since 1982 SAT scores have been equated using a different mathematical model called item-response theory (IRT). Although IRT equating methods make it possible to equate scores more accurately than linear methods when the equating samples or test forms differ in their characteristics or when new-form scores are curvilinearly related to old test forms, they are more complicated than linear methods. The latter involve only the test takers' total scores on the full and the equating tests. On the other hand, IRT equating makes use of additional information contained in the test takers' responses to individual test questions. In part, operational use of IRT equating for the SAT was dependent on the availability of high-speed computers capable of handling the substantially expanded data load and the resultant calculations. Use of IRT equating in 1982, however, was also influenced by another set of considerations.

The specifications for one of the January 1982 SAT-verbal forms called for reducing the number of very difficult items. The change was expected to produce a curvilinear relationship between scores on the new test form and scores on earlier forms. Therefore, it was considered important to use a curvilinear equating method to avoid any error that might have resulted from the use of linear equating. IRT equating not only permits curvilinear relationships, but it can also, as a true score method, adjust for a relatively large difference in equating samples. Under development for a number of years at ETS, IRT equating was used, sometimes with other methods, beginning in January 1982 to improve the accuracy of equating, particularly in the case of SAT-verbal scores.

OBSERVATIONS RELATED TO IRT EQUATING

Equating the SAT is under continuous scrutiny because the procedure is so vital to maintaining the test's respect and acceptability. However, because the new and old SAT form equating samples are not equivalent to some degree, the equating process cannot compensate completely for differences between these samples. So a persistent question is, how good was the equating? Data responsive to the question are available from samples used for the operational equatings of the test. Before 1981 these samples were selected from the total population of test takers. Since 1981 equating samples have been restricted to high school juniors and seniors and have excluded junior high, high school sophomore, and adult test takers. From an earlier discussion the reader may recall that this change coincided with a change made in selecting test analysis samples.

Equating Indices

Three types of indices have been used to evaluate equating:

1. Differences in equating test means and standard deviations between new-form and old-form equating samples. These calculations, made routinely, helped decide whether to use the Tucker or the Levine equating method.
2. Differences between the two equating lines on which the operational conversions are based, i.e., the absolute value of the difference between the scaled scores produced by the two lines at the midpoint of the score range.
3. The correlation between scores on the total test and the equating test for each new and old form equating sample.

A composite index, derived by combining the individual indices, served as the basis for the following observations.

For SAT-verbal scores:

- The index varied considerably over the 1970–85 era.
- No systematic pattern was noted.
- There was no evidence of a decrease in the index after changes in statistical specifications in 1982.

For SAT-mathematical scores:

- The index was less variable than for SAT-verbal scores.
- The trends in the index were unsystematic and unpredictable.

For both SAT-verbal and SAT-mathematical scores:

- There were some equatings with relatively low index values, but no signs of a general decrease in the index from 1970 to 1984.
- There was no decrease in the index when IRT equating was introduced for both SAT-verbal and SAT-mathematical scores in 1982.

Appropriateness of Equating Methods

Another issue is the appropriateness of operational equating methods compared to other methods. The question of whether curvilinear equating should have been used, particularly when the SAT was shortened in the fall of 1974, and also in January 1982, has been examined by comparing operational equating lines and equipercentile equating lines* for November and December administrations at the midpoint of the score range. The data here are relatively dense. Generally these comparisons found small departures of operational from equipercentile equating lines. Of 60 comparisons covering the period 1970–84, in only 9 instances did the operational lines deviate from the equipercentile lines by more than 5 raw score points at the midpoint. The maximum deviation was 7.3 raw score points for the December 1976 SAT-mathematical form.

Operational versus equipercentile equating comparisons were also made for three different intervals:

1. Between periods, i.e., equatings of newly developed shortened SAT forms (post 1974) linked to previously administered longer forms (before the fall of 1974).
2. Within periods, i.e., equating linking newly developed longer test forms to previously developed longer test forms or equatings linking newly developed shorter test forms to previously developed shorter test forms.
3. Mixed periods, i.e., equatings of test forms linked both to forms used prior to and after 1974.

*Equipercentile equating was routinely performed in the 1970–85 era to check on the curvilinearity of the equating line.

Overall, that is, for the entire 1970–84 era, mean differences in the equating comparisons were less than 3 raw score points. The December SAT-mathematical administrations were an exception, with a mean difference of 3.03. Moreover, the results for between, within, and mixed period comparisons were inconsistent with the kinds of differences one would expect if equipercentile equating was appropriate in 1974. Even if linear equating was inappropriate in a few instances, the effects were small and inconsequential.

A more detailed analysis of appropriateness of equating methods compared operational and experimental conversions for SAT-verbal and SAT-mathematical scores under different sets of real circumstances at different times. The operational conversions were linear for November and December 1974 and IRT for the two SAT forms used in January 1982, each of which required its own conversion lines.

The November and December 1974 SAT-verbal and SAT-mathematical operational conversions had to accommodate changed test specifications. In the analysis, the operational linear conversions were compared to experimental equipercentile conversions. No change in test specifications was involved in either of the two January 1982 SAT-mathematical forms. The operational equatings were IRT, and the experimental equatings were linear. One of the January 1982 SAT-verbal forms had no change in test specifications, but there were test specification changes for the second form. Thus eight sets of operational and experimental equating comparisons were made under circumstances of changed and unchanged specifications as follows:

1. November 1974 SAT-verbal (with changed test specifications), operational linear versus experimental equipercentile equatings.
2. November 1974 SAT-mathematical (with changed test specifications), operational linear versus experimental equipercentile equatings.
3. December 1974 SAT-verbal (with changed test specifications), operational linear versus experimental equipercentile equatings.
4. December 1974 SAT-mathematical (with changed test specifications), operational linear versus experimental equipercentile equatings.
5. January 1982 SAT-mathematical form 1, operational IRT versus experimental linear equatings.
6. January 1982 SAT-mathematical form 2, operational IRT versus experimental linear equatings.
7. January 1982 SAT-verbal form 1, operational IRT versus experimental linear equatings.
8. January 1982 SAT-verbal form 2 (with changed test specifications), operational IRT versus experimental linear equatings.

In contrast to the earlier discussion of the appropriateness of equating methods, which focused on scores at the scale midpoint, this analysis concerned score conversions at five points on the raw score scale. The major findings are summarized as follows:

- For the 1974 SAT-verbal and SAT-mathematical forms built to changed specifications, the operational linear equating results were similar to the experimental equipercentile equatings.
- The differences between the operational IRT and experimental linear equatings were smaller for the January 1982 SAT-verbal form built to changed test specifications than they were for the form built to unchanged specifications.
- The two January 1982 SAT-mathematical forms, built to unchanged test specifications, had operational-experimental equating correlations of .998 and .999. Likewise, for the two January 1982 SAT-verbal forms, one built to unchanged and the other to changed test specifications, the operational IRT and experimental linear equating correlations were .998 and .999, respectively.
- For one of the January 1982 SAT-mathematical forms, the standard deviation of the IRT equating was nearly four points higher than it was for the experimental linear equating. Although assembled to unchanged test specifications, this form yielded higher scaled scores at the upper end of the raw score scale and lower scaled scores at the lower end of the raw score scale.
- The study concluded that the January 1982 scores derived from the experimental linear equatings were similar to those resulting from IRT equating.

Scale Stability

Further evidence of the integrity of equating is found in scale stability studies. One recent study (McHale and Ninneman 1990) is pertinent because it covers the years 1973 to 1984 and thus overlaps the period of this report. The authors concluded that the SAT-verbal scale was relatively stable. For the SAT-mathematical scale, inconsistent results were produced by the study designs used. In one instance results showed an upward drift of 6 to 13 points, whereas a second study indicated a downward drift of 6 to 14 points. In either case, however, the drift was on the order of 1½ scaled score points a year.

Results

In the 1970–85 era, linear equating methods were used until 1982, when IRT (curvilinear) equating was introduced because of changes in SAT-verbal specifications. Subsequently, numerous special analyses have compared the results of IRT, linear, and equipercentile methods of equating new SAT forms under circumstances of changed and unchanged test specifications, using a variety of indices to

evaluate equating and to examine differences of equating lines at the middle as well as over the full score range. Collectively, these studies, together with SAT scale stability studies, indicate that only small or unsystematic variations in scores were produced by the different equating methods.

SUMMARY

This report has portrayed the Scholastic Aptitude Test as a dynamic instrument for measuring the verbal and mathematical development of students seeking education beyond high school. It is a key element in a program of service sponsored by the College Board for its school and college member and nonmember institutions. Although the SAT has had, and continues to have, a distinct identity, it is an instrument that has undergone some modification throughout its 66 years of existence.

The focus here has been on major modifications, four in number, made during the 1970–85 era. Each modification has been of quite different character: one change in the test itself, two changes in policy that have had an impact on the test, and one internal operational procedure change.

It is of interest to note that these four significant modifications of the SAT during the 1970–85 era have been accompanied by a basic stability and continuity of the test. There has been a modification in the length of the SAT with concomitant changes in test specifications, item types, content coverage, length of reading passages, arrangement of materials in test booklets, addition of two verbal subscores, and so on. There has been a dramatic change in policy regarding postadministration test security together with a substantial increase in new test form development and other adjustments. During the last five years of the 1970–85 era deliberate efforts were made to eliminate language and test question content from the SAT deemed offensive, inappropriate, or negative to major subgroups of test takers. More recently, traditional linear equating of test forms has been supplemented by the use of a more complex curvilinear method. Yet with all these modifications, the measurement properties of the SAT have changed little, if at all. The SAT is the same as it has been, but it is also quite different!

REFERENCES

Angoff, W. H., ed. 1971. *The College Board Admissions Testing Program: A Technical Report on Research and Development Activities Relating to the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

College Board. 1984. *Taking the SAT: A Guide to the Scholastic Aptitude Test and Test of Standard Written English*. New York: College Entrance Examination Board.

College Board. 1988. *10 SATs*, 3d ed. New York: College Entrance Examination Board.

Cruise, P. I., and E. W. Kimmel. 1990. *Changes in the SAT-verbal: A Study of Trends in Content and Gender References 1961–1987*. College Board Report 89–10. New York: College Entrance Examination Board.

Donlon, T. F., ed. 1984. *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

Dressel, P. L. 1940. "Some Remarks on the Kuder-Richardson Reliability Coefficient." *Psychometrica* 5(4): 305–310.

Educational Testing Service. 1986. *ETS Sensitivity Review Process: Guidelines and Procedures*. Princeton, N.J.: Educational Testing Service.

Educational Testing Service. 1987. *ETS Sensitivity Review Process: An Overview*. Princeton, N.J.: Educational Testing Service.

Marco, G. L., C. R. Crone, J. S. Braswell, W. E. Curley, and N. K. Wright. 1990. *Trends in SAT Content and Statistical Characteristics and Their Relationship to SAT Predictive Validity*. RR-90-12. Princeton, N.J.: Educational Testing Service.

McHale, F. J., and A. M. Ninneman. 1990. *The Stability of the Score Scale for the Scholastic Aptitude Test from 1973–1984*. RR-90-6. Princeton, N.J.: Educational Testing Service.

APPENDIX: SAT ITEM-TYPE EXAMPLES

SAT Directions and Sample Questions

SAT-VERBAL

Antonyms, Analogies, Sentence Completions, Reading Comprehension

SECTION **1** Time—30 minutes 40 Questions

For each question in this section, choose the best answer and fill in the corresponding oval on the answer sheet.

Antonyms

Each question below consists of a word in capital letters, followed by five lettered words or phrases. Choose the word or phrase that is most nearly opposite in meaning to the word in capital letters. Since some of the questions require you to distinguish fine shades of meaning, consider all the choices before deciding which is best.

Example:

GOOD: (A) sour (B) bad (C) red (D) hot (E) ugly

Ⓐ ● Ⓒ Ⓓ Ⓔ

Sample Questions

1. SURPLUS: (A) shortage (B) criticism (C) heated argument (D) sudden victory (E) thorough review

2. TEMPESTUOUS: (A) responsible (B) predictable (C) tranquil (D) prodigious (E) tentative

Correct Answers: 1. A 2. C

Analogies

Each question below consists of a related pair of words or phrases, followed by five lettered pairs of words or phrases. Select the lettered pair that best expresses a relationship similar to that expressed in the original pair.

Example:

YAWN : BOREDOM :: (A) dream : sleep (B) anger : madness (C) smile : amusement (D) face : expression (E) impatience : rebellion

Ⓐ Ⓑ ● Ⓓ Ⓔ

Sample Questions

3. APPAREL : SHIRT :: (A) sheep : wool (B) foot : shoe (C) light : camera (D) belt : buckle (E) jewelry : ring

4. BUNGLER : SKILL :: (A) fool : amusement (B) critic : error (C) daredevil : caution (D) braggart : confidence (E) genius : intelligence

Correct Answers: 3. E 4. C

Sentence Completions

Each sentence below has one or two blanks, each blank indicating that something has been omitted. Beneath the sentence are five lettered words or sets of words. Choose the word or set of words that, when inserted in the sentence, best fits the meaning of the sentence as a whole.

Example:

Although its publicity has been ----, the film itself is intelligent, well-acted, handsomely produced, and altogether ----.

(A) tasteless. . .respectable (B) extensive. . .moderate (C) sophisticated. . .amateur (D) risqué. . .crude (E) perfect. . .spectacular

● Ⓑ Ⓒ Ⓓ Ⓔ

Sample Questions

5. Either the sunsets at Nome are ----, or the one I saw was a poor example.

(A) gorgeous (B) overrated (C) unobserved (D) exemplary (E) unappreciated

6. Specialization has been emphasized to such a degree that some students — nothing that is — to their primary area of interest.

(A) ignore. . .contradictory
(B) incorporate. . .necessary
(C) recognize. . .fundamental
(D) accept. . .relevant
(E) value. . .extraneous

Correct Answers: 5. B 6. E

Each passage below is followed by questions based on its content. Answer the questions following each passage on the basis of what is stated or implied in that passage.

From the beginning, this trip to the high plateaus in Utah has had the feel of a last visit. We are getting beyond the age when we can unroll our sleeping bags under any pine or in any wash, and the gasoline situation throws the future of automobile touring into doubt. I would hate to have missed the extravagant personal liberty that wheels and cheap gasoline gave us, but I will not mourn its passing. It was part of our time of wastefulness and excess. Increasingly, we will have to earn our admission to this spectacular country. We will have to come by bus, as foreign tourists do, and at the end of the bus line use our legs. And if that reduces the number of people who benefit every year, the benefit will be qualitatively greater, for what most recommends the plateaus and their intervening deserts is not people, but space, emptiness, silence, awe.

I could make a suggestion to the road builders, too. The experience of driving the Aquarius Plateau on pavement is nothing like so satisfying as the old experience of driving it on rocky, rutted, chuckholed, ten-mile-an-hour dirt. The road will be a lesser thing when it is paved all the way, and so will the road over the Fish Lake Hightop, and the one over the Wasatch Plateau, and the steep road over the Tushar, the highest of the plateaus, which we will travel tomorrow. To substitute comfort and ease for real experience is too American a habit to last. It is when we feel the earth rough to all our length, as in Robert Frost's poem, that we know it as its creatures ought to know it.

The reading passages in this test are brief excerpts or adaptations of excerpts from published material. The ideas contained in them do not necessarily represent the opinions of the College Board or Educational Testing Service. To make the text suitable for testing purposes, we may in some cases have altered the style, contents, or point of view of the original.

7. According to the author, what will happen if fewer people visit the high country each year?
   (A) The characteristic mood of the plateaus will be tragically altered.
   (B) The doctrine of personal liberty will be seriously undermined.
   (C) The pleasure of those who do go will be heightened.
   (D) The people who visit the plateaus will have to spend more for the trip.
   (E) The paving of the roads will be slowed down considerably.

8. The author most probably paraphrases part of a Robert Frost poem in order to
   (A) lament past mistakes
   (B) warn future generations
   (C) reinforce his own sentiments
   (D) show how poetry enhances civilization
   (E) emphasize the complexity of the theme

9. It can be inferred from the passage that the author regards the paving of the plateau roads as
   (A) a project that will never be completed
   (B) a conscious attempt to destroy scenic beauty
   (C) an illegal action
   (D) an inexplicable decision
   (E) an unfortunate change

Correct Answers: 7. C
8. C
9. E

BEST COPY AVAILABLE

SAT- MATHEMATICAL

Regular Mathematics, Data Sufficiency, Quantitative Comparisons

SECTION 2 Time—30 minutes, 25 Questions

In this section solve each problem, using any available space on the page for scratchwork. Then decide which is the best of the choices given and fill in the corresponding oval on the answer sheet.

The following information is for your reference in solving some of the problems.

Circle of radius r: Area $= \pi r^2$; Circumference $= 2\pi r$
The number of degrees of arc in a circle is 360.
The measure in degrees of a straight angle is 180.

Definition of symbols:

$=$ is equal to	$\leq$ is less than or equal to
$\neq$ is unequal to	$\geq$ is greater than or equal to
$<$ is less than	$\parallel$ is parallel to
$>$ is greater than	$\perp$ is perpendicular to

Triangle: The sum of the measures in degrees of the angles of a triangle is 180.

If $\angle CDA$ is a right angle, then

(1) area of $\triangle ABC = \frac{AB \times CD}{2}$

(2) $AC^2 = AD^2 + DC^2$

Note: Figures that accompany problems in this test are intended to provide information useful in solving the problems. They are drawn as accurately as possible EXCEPT when it is stated in a specific problem that its figure is not drawn to scale. All figures lie in a plane unless otherwise indicated. All numbers used are real numbers.

Regular Mathematics

Sample Questions

1. If $2y = 3$, then $3(2y)^2 =$

(A) $\frac{27}{4}$

(B) 18

(C) $\frac{81}{4}$

(D) 27

(E) 81

2. Of seven consecutive integers in increasing order, if the sum of the first three integers is 33, what is the sum of the last three integers?

(A) 36
(B) 39
(C) 42
(D) 45
(E) 48

Correct Answers: 1. D
2. D

Directions: Each of the data sufficiency problems below consists of a question and two statements, labeled (1) and (2), in which certain data are given. You have to decide whether the data given in the statements are sufficient for answering the question. Using the data given in the statements plus your knowledge of mathematics and everyday facts (such as the number of days in July or the meaning of counterclockwise), you are to fill in the corresponding oval

- A if statement (1) ALONE is sufficient, but statement (2) alone is not sufficient to answer the question asked;
- B if statement (2) ALONE is sufficient, but statement (1) alone is not sufficient to answer the question asked;
- C if BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient;
- D if EACH statement ALONE is sufficient to answer the question asked;
- E if statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

Numbers: All numbers used are real numbers.

Figures: A figure in a data sufficiency problem will conform to the information given in the question, but will not necessarily conform to the additional information given in statements (1) and (2).

You may assume that lines shown as straight are straight and that angle measures are greater than zero.

You may assume that the position of points, angles, regions, etc., exist in the order shown.

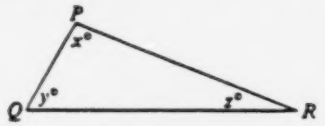
All figures lie in a plane unless otherwise indicated.

Example:

In $\triangle PQR$, what is the value of x?

(1) $PQ = PR$

(2) $y = 40$



Explanation: According to statement (1), $PQ = PR$; therefore, $\triangle PQR$ is isosceles and $y = z$. Since $x + y + z = 180$, $x + 2y = 180$. Since statement (1) does not give a value for y, you cannot answer the question using statement (1) by itself. According to statement (2), $y = 40$; therefore, $x + z = 140$. Since statement (2) does not give a value for z, you cannot answer the question using statement (2) by itself. Using both statements together, you can find y and z; therefore, you can find x, and the answer to the problem is C.

Sample Questions

3. Is $a + b = a$?

   (1) $b = 0$

   (2) $a = 10$

4. Is rectangle R a square?

   (1) The area of R is 16.

   (2) The length of a side of R is 4.

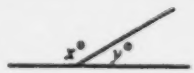
Correct Answers: 3. A
4. C

Questions 5-6 each consist of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet fill in oval

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined from the information given.

AN E RESPONSE WILL NOT BE SCORED.

EXAMPLES

	Column A	Column B	Answers
E1.	2×6	$2 + 6$	● Ⓑ Ⓒ Ⓓ Ⓔ
E2.	$180 - x$ 	y	Ⓐ Ⓑ ● Ⓓ Ⓔ
E3.	$p - q$	$q - p$	Ⓐ Ⓑ Ⓒ ● Ⓔ

Notes:

1. In certain questions, information concerning one or both of the quantities to be compared is centered above the two columns.
2. In a given question, a symbol that appears in both columns represents the same thing in Column A as it does in Column B.
3. Letters such as x, n, and k stand for real numbers.

Sample Questions

	Column A	Column B
5.	The least positive integer divisible by 2, 3, and 4	24

Parallel lines ℓ_1 and ℓ_2 are 2 inches apart. P is a point on ℓ_1 and Q is a point on ℓ_2.

6.	Length of PQ	3 inches

Correct Answers: 5. B
6. D

Examples of Explained SAT Items*

Analogy Example

Remember that a pair of words can have more than one relationship. For example:

PRIDE : LION : : (A) snake : python (B) pack : wolf
(C) rat : mouse (D) bird : starling (E) dog : canine

A possible relationship between *pride* and *lion* might be that "the first word describes a characteristic of the second (especially in mythology)." Using this reasoning, you might look for an answer such as *wisdom* : *owl*, but none of the given choices has that kind of relationship. Another relationship between *pride* and *lion* is "a group of lions is called a pride"; therefore, the answer is (B) *pack* : *wolf*; "a group of wolves is called a pack."

Mathematics Example

If $16 \cdot 16 \cdot 16 = 8 \cdot 8 \cdot P$, then P =

(A) 4 (B) 8 (C) 32 (D) 48 (E) 64

This question can be solved by several methods. A time-consuming method would be to multiply the three 16s and then divide the result by the product of 8 and 8. A quicker approach would be to find what additional factors are needed on the right side of the equation to match those on the left side. These additional factors are two 2s and a 16, the product of which is 64. Yet another method involves solving for P as follows:

$$P = \frac{\cancel{16}^{2} \cdot \cancel{16}^{2} \cdot 16}{\cancel{8} \cdot \cancel{8}} = 2 \cdot 2 \cdot 16 = 64$$

The correct answer is (E).

*From *Taking the SAT* (College Entrance Examination Board, 1984).

Sample Minority-Relevant Reading Passage and Reading Comprehension Test Items

In certain non-Western societies, a scholar once suggested, the institution of communal music "gives to individuals a solid center in an existence that seems to be almost chaos, and a continuity in their being that would otherwise too easily dissolve before the calls of the implacable present. Through its words, people who might be tempted to give in to the malice of circumstances find their old powers revived or new powers stirring in them, and through these life itself is sustained." This, I think, sums up the role played by song in the lives of Black American slaves. Songs of the years before Emancipation supply abundant evidence that in the structure of the music, in the survival of oral tradition, and in the ways the slaves expressed their Christianity, important elements of their common African heritage became vitally creative aspects of their Black American culture.

Although it was once thought that Africans newly arrived in America were passive recipients of European cultural values that supplanted their own, most African slaves in fact were so isolated from the larger American society beyond the plantation that European cultural influence was no more than partial. Through necessity they drew on and reestablished the only cultural frame of reference that made any sense to them, that of cultures in which they had been raised, and thus passed on distinctly African cultural patterns to slaves born and raised in America. One example of the process of cultural adjustment is the response to Christianity, valued by American-born slaves not as an institution imported intact from another culture but as a spiritual perception of heroic figures and demonstration of divine justice that they could use to transcend the bonds of their condition through the culturally significant medium of song. Earlier historians frequently failed to perceive the full importance of this because they did not take seriously enough the strength of feeling represented in the sacred songs. A religion that was a mere anodyne imposed by an oppressing culture could never have inspired a music as forceful and striking to all who witnessed it as the religious music of the American slaves. Historians who have tried to argue that these people did not oppose the institution of slavery in any meaningful collective way reason from a narrowly twentieth-century Western viewpoint. Within the frame of reference inherited from African cultures in which the functions of song included criticism and mockery of rulers, there were meaningful methods of opposition to authority and self-assertion so foreign to most Western historians as to be unrecognizable. Modern Americans raised in a wholly Western culture need to move mentally outside their own culture, in which music plays only a peripheral role, before they can understand how American-born slaves put to use the functions music had had for their African ancestors.

31. Which of the following statements best expresses the main idea of the passage?

(A) Communal song was a vital part of the heritage of American-born slaves.
(B) Communal music was primarily important as a recreational pastime.
(C) Communal song had several functions in ancient African cultures.
(D) Non-Western cultures give Western historians insights into elements of their own culture.
(E) Songs prized by American slaves reminded them of their legendary homelands.

32. Underlying the scholar's description of communal music (lines 2-9) is the assumption that human existence

(A) presents a rich variety of subjects for popular songs
(B) sometimes reveals profound artistic truths
(C) often seems bewildering and hopeless
(D) crushes the spirit of even the most resourceful artist
(E) rarely encourages artistic expression

33. The passage suggests that the "strength of feeling" (line 37) expressed in the slaves' religious music was a direct reflection of which of the following?

(A) The use of religion to distract people from present problems
(B) The sense of injustice as inevitable and inescapable
(C) The perception of Christianity as a distinctly foreign institution
(D) The role of song as a vehicle for social commentary
(E) The importance of music to Western society

GO ON TO THE NEXT PAGE

34. The sentence that begins "A religion that was . . ." (lines 38-42) makes all of the following points EXCEPT:

(A) Christianity had important spiritual significance for Black American slaves.
(B) People who have assumed that slaves considered Christianity only a superficial, alien institution are mistaken.
(C) The power of the slaves' religious music indicates how deeply they felt what they sang.
(D) The slaves' religious music provided a moving experience for the listeners as well as the singers.
(E) The religious music of American-born slaves made only indirect use of Christian figures.

35. The author states that historians have sometimes misunderstood the music of American slaves for which of the following reasons?

I. They could not personally observe the music being sung by American slaves.
II. Some historians interpreted the significance of this music within the wrong cultural context.
III. Many Western historians felt uncomfortable with the presentation of religious stories through art.

(A) I only
(B) II only
(C) III only
(D) I and II
(E) I and III